Event Detection based on Dockless Scooter Trips in Austin, Texas

Soham Mody (srm699@nyu.edu)* Timur Mukhtarov (tm1722@nyu.edu)*

*MS students, Center for Urban Science and Progress, New York University

Abstract

Dockless scooters are a new addition to ride sharing networks of cities. Despite their convenience, there are some drawbacks associated with them. One problem is that these scooters are often left on the streets, blocking both private and public property. This problem is exacerbated by large events that often take place in cities. In this study, we explore and understand Austin, TX dockless scooter ridership in a data-driven fashion. We develop three event detection models using an univariate forecasting model as well as machine learning ensemble methods. The developed methodology can be used by the city to for a more optimal resource allocation and ultimately to prevent sidewalks from being cluttered with scooters.

1. Introduction

Ride sharing is a recent but widely accepted addition to transportation networks of cities around the world. Among the newest forms of shared-use mobility, dockless electric scooters have gained an almost instant popularity in various cities across the United States and globally. Some figures suggest that there are about 15 companies operating dockless scooters in about 30 states and 130 cities (Smart City Dive, 2019)

In addition to the obvious advantages like being readily available and affordable, the scooters have unique characteristics compared to other recently introduced forms of shared-use transportation. First, they differ from bike-sharing systems in that they lack docks and thus are often spread out all around cities. Secondly, there are few limits to how many scooters can be located in a certain areas, allowing companies to experiment with recalibration and users to park the scooters right outside their final destination. These unique characteristics, however, bring some challenges for cities and local problems. One significant problem is that riders leave the scooters everywhere thus impairing the orderliness and safety of streets and sidewalks.

This problem leads to cities having to spend more resources on some of the city operations. The cities often have to deploy people to move scooters intruding on private property. Further, it disturbs the lives of people living in the cities because they feel scared of being knocked over by amateur riders, the scooters might block their private property or the accessibility infrastructure. Public safety is also another concern. Recently, the Center for Disease Control (CDC) had conducted a study to understand how scooter injuries occur and what the risk and protective factors for injuries might be (CDC, 2018). The study investigated 936,110 trips taken on the dockless e-scooters and found 192 people injured via these trips who had to make emergency-room visits.

The issue gets even more out-of-control when masses gather for an event. This had happened in Austin during the SXSW festival, which brings between 200,000 to 400,000 attendees to the city each year. Festival-goers gladly welcomed the scooters after being frustrated with having to travel long distances between events in previous years. Scooters offered a fun and cheap way to travel to conference events. Thousands of scooters could be found in downtown Austin, many of them obstructing the sidewalks,

roads, and private spaces. Riders were mostly not wearing helmets while cruising at high speeds (Goard, 2019).

In this study, we take a data-driven approach to explore and understand how electronic dockless scooter ridership can be used to detect significant social events, concerts, major holidays or extreme weather in Austin, Texas.

Based on these questions, we formulated a few goals to guide our research process:

1. Understand and explore origins and destination of rides on a census tract level.
2. Formulate a baseline model that uses a number of rides as a predictor.
3. Formulate a more sophisticated predictive model that uses further features like trip duration and distance travelled.
4. Explore both traditional time-series analysis methods like ARMA and other anomaly detection methods like Isolation Forest and Local Outlier Factor.

2. Related Works

There are no published studies done on dockless scooters. However, various studies have been done on other forms of shared-use mobility. Particularly, some studies have focused on forecasting events using bike-sharing systems like Citi Bike and ride-sharing services like Uber and Lyft.

For predicting extreme event forecasting using ride-sharing data, Laptev at al. propose using an end-to-end recurrent neural network architecture (Laptev at al., 2017). Their Long Short Term Memory (LSTM) implementation for time-series analysis outperforms other forecasting methods such as quantile random forest on Uber data. Their model generalizes well to a well-known M3-competition dataset created by the International Institute of Forecasters.

A couple of studies have focused on bike-sharing data. Yang et al. researched mobility modeling and prediction in Hangzhou, the world's largest public bike-sharing system (Yang et al., 2016). They developed a spatio-temporal bicycle mobility model based on historical bike-sharing data, and devised a traffic prediction mechanism on a per-station basis with sub-hour granularity. Their prediction model is based on random forest model, and it outperforms the benchmark Historical Average (HA) and Autoregressive and Moving Average (ARMA) models. Li et al. developed methods to predict ride-sharing bicycle traffic amount between clusters of stations (Li et al., 2015). They predict the total number of bikes that will be rented out in New York and Washington, DC using a gradient boosting regression tree.

3. Data

Datasets

Our main dataset is of Dockless Vehicle Trips originally from the open data portal of the City of Austin. This dataset contains dockless vehicle trip data reported to the City of Austin Transportation Department as part of the Dockless Mobility Service operating rules.

We analyzed an archived version of this dataset. It consists of about 2,746,505 dockless vehicle rides collected in Austin, TX from April 2018 to February 2019. Each record consists of the following fields: device ID, trip duration, vehicle type (scooter or bike), trip distance, start time, end time, date, day of week, council district (start and end), census tract (start and end), latitude, and longitude.

A more detailed description of this and other datasets can be found in Appendix 1.

Pre-processing and feature engineering

Before conducting actual analyses and modeling, we needed to clean and prepare the data. There were trips where start or end locations were missing and instead in some cases, an indicator for the same called 'Out-of-bounds' was present so, we removed all the rows either missing locations or containing 'Out_of_bounds'. Then, we also removed rides with dockless bicycle as vehicle type as there was very less data about them and they were not the focus of our project. At this point, we had 2,612,879 records in our dataset, which about a 5 percent reduction from our original dataset.

We also obtained a shapefile for Austin's census tracts and performed a spatial merge with our rides data based on start location in that. We have not used the census tracts in our actual model, but they were useful in getting an idea of flows of the trips through the city during exploratory analysis.

Another major issue we ran into and didn't expect was that the dataset contained trips with negative trip duration and distances along with highly unreasonable values of these variables like 1 million kilometers and so on. So, we decided to retain only the trips with duration and distance between 10 and 40,000. The motivation for choosing 40,000 meters was that dockless scooters have a maximum capacity of about 25 miles before it needs to be recharged. This resulted in the loss of almost 10% of our dataset.

For our models, we needed to aggregate various features by day to produce a time series. These included the total number of daily trips, average distance covered in a trip per day and average duration of a trip per day.

In the end our dataset looked like the following table:

Date	Number of Trips	Avg distance	Avg Duration
2018-05-23	318	2172.723270	897.798742
...	...	...	...
2019-02-12	300	1296.960000	552.820000

Table 1. Aggregated dataframe

Exploratory Data Analysis

To explore our dataset, we first plotted a time series chart of all dates. As you can see from **Figure 1**(left), directly plotting the time series makes it difficult to understand the plot but smoothing the time series by applying a rolling mean makes it easier to visualize and interpret(right). The plots report the daily number of rides for all scooters in Austin Texas from June 2018 until February 2019. We have filtered out the previous trips till May as they were either very few trips or they were not continuous and had missing weeks in between.

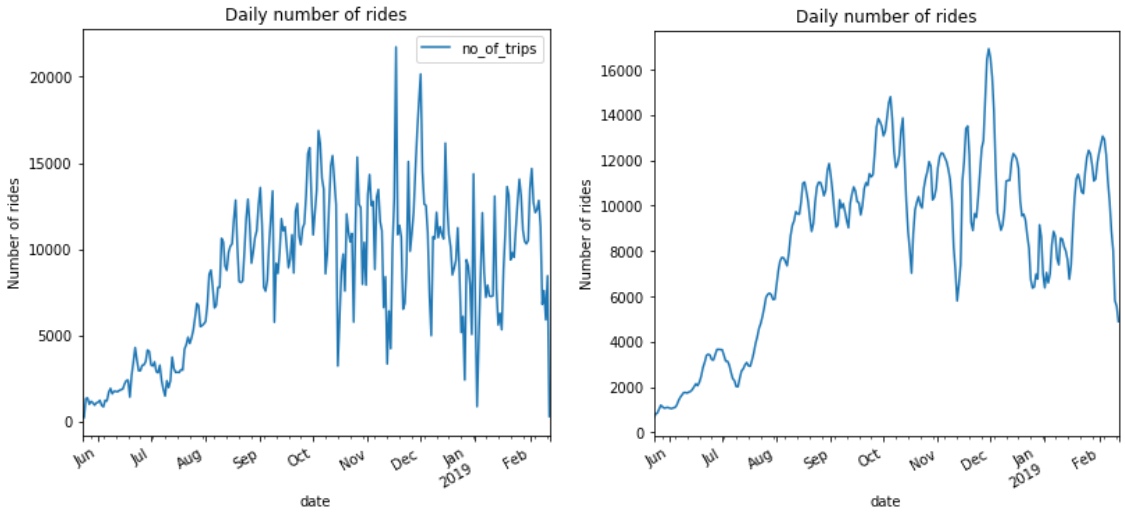

Figure 1. Daily number of rides before(left) and after smoothing(right)

We can see an unusual dip in the number of rides till July 2018 and that the trips are becoming more prevalent only around August 2018. After searching this on the internet, we realized that in August 2018, the scooter companies expanded their operations and hence, had more scooters on the streets.

Then, we looked into the census tracts and geographical data to see where the trips were being originated and concluded. Figure 2 illustrates the location of origin points of rides. In the graph, each point represents a geographic location of the origin of the trip. Each point has an opacity of 50%, so darker regions on map denote areas with more rides.

From the graph, we saw that trips mostly originate in Downtown Austin and adjacent neighborhoods such as Clarksville, and Central East Austin. This makes sense as these are commercial areas with a lot of offices and hotels. A lot of rides also originate in residential neighborhoods both close to downtown (Travis Heights) and somewhat farther from it (Hyde Park). The situations is similar with drop offs, except we see an even higher activity around downtown and adjacent areas. We can also see that a lot of neighborhoods in Austin don't have any scooters. A possible explanation for that is that these areas consist of mostly single family house that are more car-dependent.

Figure 2. Scooter rides origins

We looked at time patterns of ridership to understand when people use the dockless scooters. In Figure 4 we can see that scooters are popular throughout the day. While there is a pick up in the morning, the fact that the ridership is consistently high throughout the afternoon suggests that scooters are used for leisurely activities such as sightseeing rather than for commuting. From this, it is possible that more rides take place on weekends rather than on weekdays. Thus, there might be some sort of weekday seasonality, which we considered while choosing and tweaking a time-series forecasting modeling technique.

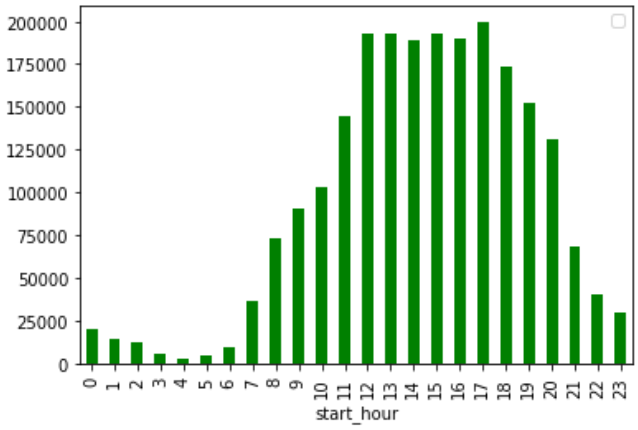

Figure 3. Number of trips by hour of the day

In preparation for modeling an univariate time-series forecasting model, we also had to test our data for autocorrelation. As seen in Figure 4, the autocorrelation plot tells us that each observation has a strong positive correlation with its previous 50 observations which means we can apply an Autoregressive(AR) model.

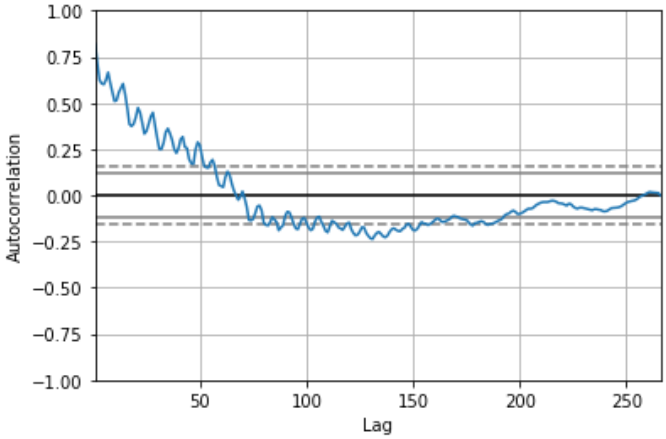

Figure 4. The autocorrelation plot for the time series

4. Methods

After gaining a better understanding of our data, we proceeded with creating a model to predict events in Austin using dockless scooter data. Identifying big events happening in the city is essentially an anomaly detection problem since such events don't happen as often. Based on that and on our initial goals, we used a common univariate time series forecasting model called Seasonal Autoregressive Integrated Moving Average (SARIMA) as our baseline forecasting model, and then proceeded with other more complicated anomaly detection methods like isolation forests and local outlier factor (LOF) methods.

Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA is a modification on the most popular forecasting methods for univariate time series data forecasting which is, Autoregressive Integrated Moving Average (ARIMA). ARIMA model supports both an autoregressive and moving average elements of the time-series. The integrated element refers to differencing allowing the method to support time series data with a trend. SARIMA in addition to supporting these elements, also explicitly supports univariate time series data with a seasonal component.

We decided that the seasonal component is important since we saw the weekday seasonality pattern earlier in our analysis. The seasonal part of the model consists of terms that are very similar to the non-seasonal components of the model, but they involve backshifts of the seasonal period. (Hyndman & Athansapoulos, 2013). So, we applied a SARIMA model of the order (1,1,1) with a seasonality of 7 (1 week) on the time-series data containing the daily number of trips.

We performed a train-test split keeping 90 observations in the test set. We also applied logarithmic transformation to the data as a pre-processing step for the model. We trained the SARIMA on the first 176 days and then forecasted the results for the remaining 90. Then, based on the differences between the actual and the predicted values, we decided to declare observations where the error in the prediction was more than 2 standard deviations away from the mean as anomalous.

Isolation Forest

The algorithm is based on the fact that anomalies are data points that are few and different (Liu et al., 2008). We decided to use this method due to its speed and scalability. We found these qualities beneficial due to the size of our initial dataset.

This method works differently than many other anomaly detection models. Usually, anomaly detection algorithms construct a profile of normal instances, then identify instances that do not conform to the normal profile as anomalies. This method explicitly isolates each point in the data and splits them into outliers or inliers. It starts with multiple trees in a totally random forest. It splits data depending on how long it takes to separate the points. At each point, it randomly selects split attribute and value. Then, it measures the average number of splits needed to isolate each data point from the rest of the data.

We employed this method knowing that it is a linear-type algorithm and its decision boundaries are vertical or horizontal. As these decision lines can only be parallel to the axes, there is a possibility that there are regions that contain many branch cuts and only a few or single observations. Thus, this model may results in improper anomaly scores for some of the observations.

We used a multivariate time series for this model containing the total number of daily trips, average distance covered in a trip per day and average duration of a trip per day. We standardized all 3 features and tried various values for the contamination(percentage of total observations which are outliers) parameter in order to see if the predicted anomalies make sense. For that, we made a 3-D scatter plot of the 3 features as well as applied PCA on the 3 standardized features and then, made a scatter plot of the 2 most prominent features. There is a obvious trade-off here between the number of False Positives due to higher contamination and number of False Negatives due to lower contamination which would result in anomalies being missed. In our case, it would be better to identify some normal days as anomalous rather than missing anomalous days and classifying them as normal.

Local outlier factor (LOF)

We also wanted to use a density-based outlier detection model so, local outlier factor model seemed like the way to go forward. Our reasoning behind that is that these models are able to identify local outliers in a data set better than models that consider outliers only with respect to the global data distribution.

Local density-based models perform density estimation at each data point, and identify records in low-density regions as potential anomalies. LOF compares each point's density to the density of it's n-nearest neighbours (Breunig et al., 2000). It then reports the points with the lowest ratios as outliers.

While using this method, we were aware of its limitations when it comes to datasets that are large and that have a very high representational dimensionality. It is known that computations associated with this method are memory-intensive. In this research, however, the usage of this model is justified as our final dataset is of a manageable size and has only 3 features that need to be modelled.

Just like Isolation Forest, we apply PCA on the 3 standardized features for LOF and then we make a plot of the 2 most prominent features to decide an ideal contamination value and number of nearest neighbours. To show which points are more anomalous, there is the negative outlier factor(nof) score which is close to -1 for inliers and more negative for outliers. So, the plot also plots circles around the points proportional to their nof score.

4 Results

Seasonal Autoregressive Integrated Moving Average (SARIMA)

In Figure 5, we can see that the model detected five anomalies. And they seem to make sense as they have either exceptionally high or low number of trips (Table 2). For example, December 31st has more than 14,000 which is very likely as many tourists might have come to Austin for New Year's Eve. Also, February 12th has just 300 trips and is classified an anomaly because that is the last date in our data and data collection was most likely stopped very early in the day.

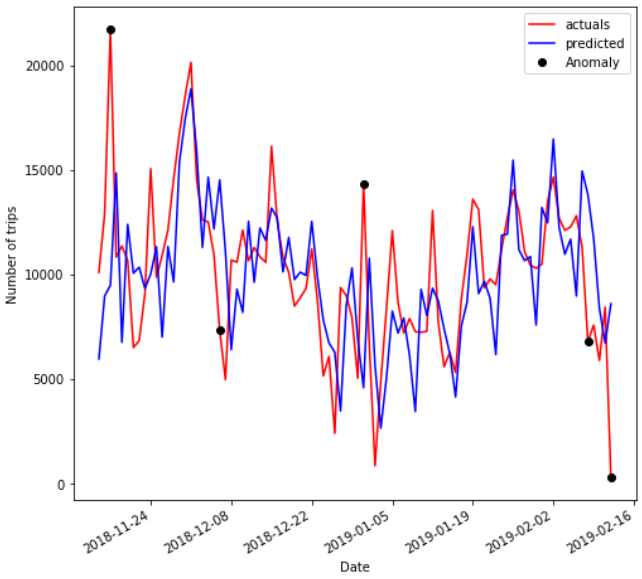

Figure 5. SARIMA model

	date	actuals	predicted	error	percentage_change	meanval	deviation	anomaly
2	2018-11-17	21727	9505.130453	12221.869547	56.251989	229.24923	3339.326446	1
21	2018-12-06	7343	14536.223550	-7193.223550	-97.960283	229.24923	3339.326446	1
46	2018-12-31	14369	4603.836887	9765.163113	67.959935	229.24923	3339.326446	1
85	2019-02-08	6802	13841.631193	-7039.631193	-103.493549	229.24923	3339.326446	1
89	2019-02-12	300	8614.402073	-8314.402073	-2771.467358	229.24923	3339.326446	1

Table 2. SARIMA results

We can verify from both the plot and table that the anomalies actually had quite different data than what was expected that days(predictions). So, this method is accurate but, it considers only 1 feature and in that also, there is relatively less flexibility in controlling the percentage of outliers.

Isolation Forest

As seen in Figure 6, we have tried various values for the contamination. For us, 0.14 seems to be a good value for contamination as 0.08 and 0.12 miss a few anomalies and our target is to be able to detect more anomalies even at the cost of increasing the number of False positives(Uptil a certain extent).

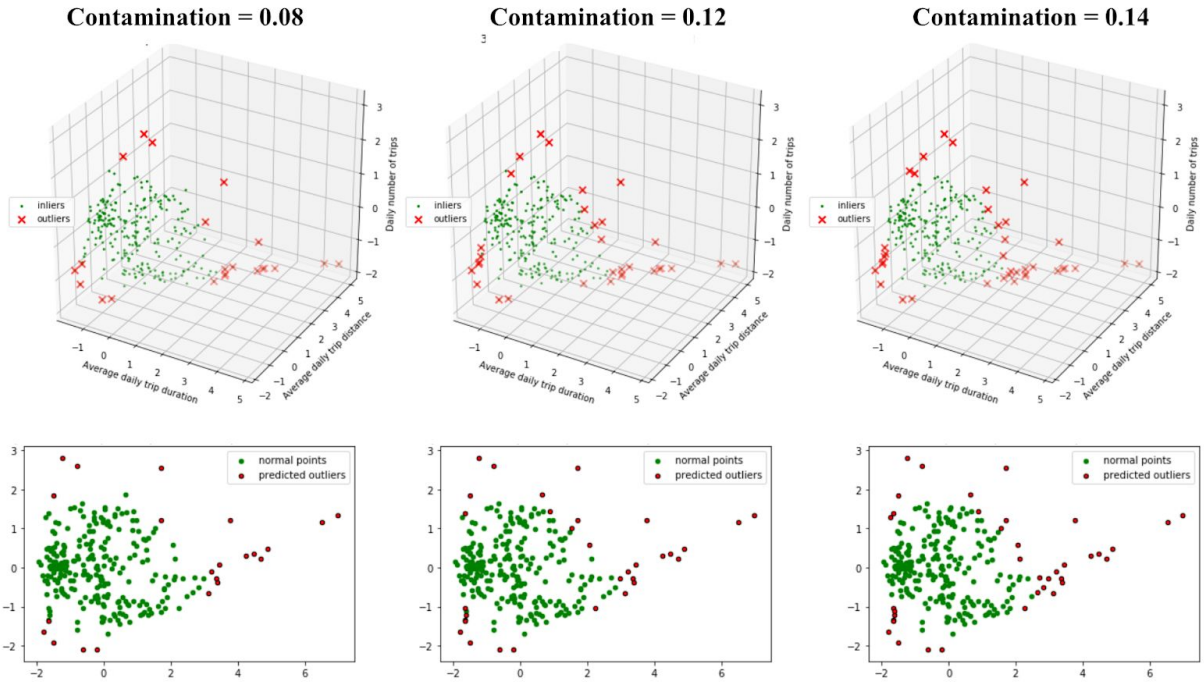

Figure 6. Isolation Forest scatterplots

We can see that there are a few points that you might feel are not outliers looking at just one of the **Figure 7** graphs, but looking at the others, you can see that they indeed have anomalous values for those. This goes to show that a values can be abnormal on just one scale and still, be an anomaly.

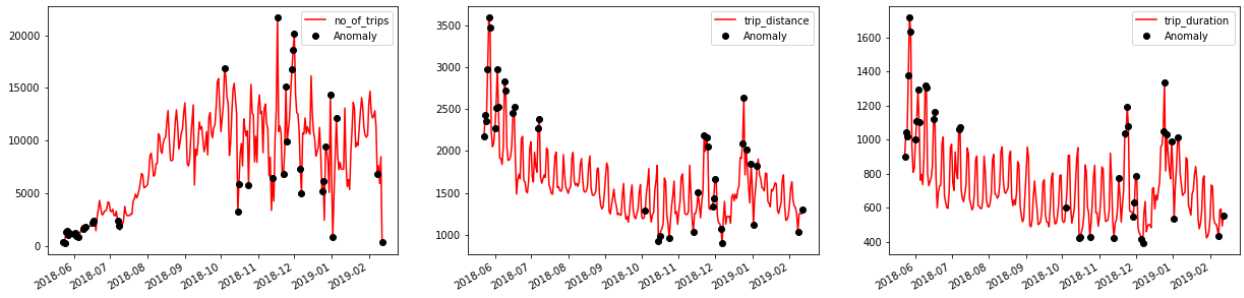

Figure 7. Isolation Forest Anomaly Detection by number of trips (left), trip distance (middle), and trip duration (right)

Local outlier factor (LOF)

As seen in Figure 8 and 9, the value for nearest neighbours is also very crucial in addition to the contamination as if that value is like 10, the entire small-valued cluster at the start of the series would be considered anomalous. So, it is important to pick a higher value here.

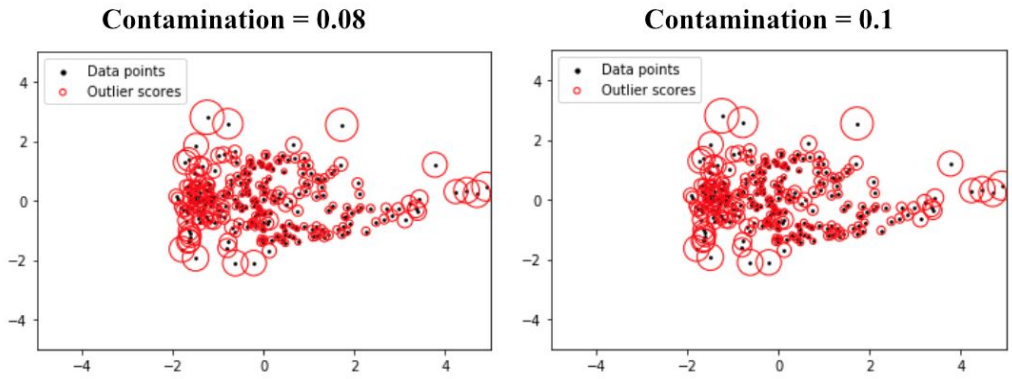

Figure 8. Local outlier factor scatterplots

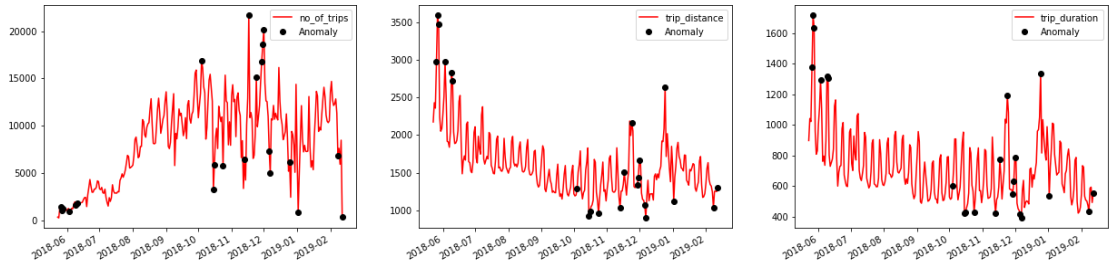

Figure 9. Local outlier factor anomaly detection by number of trips (left), trip distance (middle), and trip duration (right)

5 Conclusions and Further work

From the results, we can see that, both the Isolation Forest and the Local Outlier Factor models detect more events compared to the baseline model. This is due to the contamination being more flexible. Thus, we are able to detect more events even though this may increase the number of normal events being detected as anomalies in some cases. Additionally, we can see how Isolation Forest just sees things on a global scale and classifies events just on that basis which might result in it missing many local outliers and even more normal events being classified as anomalies. But, density-based methods like LOF solve this problem by looking at the observation from a local perspective and are able to detect these local outliers. Of course in the case of a huge city like Austin, none of these results can conclusively predict that an event did occur as there could be many external factors for the abnormal shift in the values of these features. The results would always need to be validated against real world information to see if they are correct.

We believe that our research has a few applications directly relevant to the field of urban informatics and systems. This research can be used for a more optimal resource allocation for prevention sidewalks from being cluttered with scooters. The city can use this model to predict when large-scale events and deploy the number of its employees accordingly. They can also use this model to set preventative measures such as setting up designated scooter parking spaces for events ahead of time. Another example of using this for resource allocation would be using it to make decisions about closing sidewalks or roads for maintenance. Another application is anomaly detection. The City of Austin can use this research to monitor the city for extreme events such as terrorist attacks or government shutdowns.

There is a room for improvement in future work. As the size of Austin's dockless vehicle trips will only get higher, more efficient and scalable models must be used in the future. One place to start would be to use the extended isolation forest method, which is an extension of a method used in this paper. Another promising area is performing similar analysis with recurrent neural networks.

Another area we plan to explore in the future is forecasting event locations. This research has laid some groundwork for that by exploring top origins and destinations of rides, and also by merging census tract data into the Dockless Vehicle Trips dataset.

Bibliography

Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). Lof. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data - SIGMOD 00*. doi:10.1145/342009.335388

Centers for Disease Control and Prevention. (2018, November). *Characterization of Dockless Electric Scooter Related Injury Incidents - Austin, Texas, September-November, 2018*[Press release]. Retrieved from https://www.cdc.gov/eis/conference/dpk/Dockless_Electric_Scooter_Related_Injury.html

Census Tract Fill. (2019). Retrieved from City of Austin Open Data Portal.

Dockless Vehicle Trips. (2019). Retrieved from City of Austin Open Data Portal.

Goard, A. (2019, February 27). Austin has a game plan for the first SXSW with electric scooters, bikes. *KXAN News*. Retrieved from https://www.kxan.com/news/local/austin/austin-has-a-game-plan-for-the-first-sxsw-with-electric-scooters-bikes/1815136176

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. Heathmont, Vic.: OTexts.

Laptev, N., Yosinski, J., Li, L. E., & Smyl, S. (2017). Time-series Extreme Event Forecasting with Neural Networks at Uber. *2017 International Conference on Machine Learning (ICML)*. doi:10.1109/icdmw.2017.19

Li, Y., Zheng, Y., Zhang, H., & Chen, L. (2015). Traffic prediction in a bike-sharing system. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS 15*. doi:10.1145/2820783.2820837, https://dl.acm.org/citation.cfm?doid=2820783.2820837

Liu, F. T., Ting, K. M., & Zhou, Z. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*. doi:10.1109/icdm.2008.17

Mapping the impact of dockless vehicles. (2019, May 3). *Smart Cities Dive*.

Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., & Moscibroda, T. (2016). Mobility Modeling and Prediction in Bike-Sharing Systems. *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys 16*. doi:10.1145/2906388.2906408

Code

The project was implemented in Python using pre-existing packages including, but not limited to numpy, matplotlib, and sklearn.

The code can be accessed in the Jupyter Notebook format on the project's GitHub page at https://github.com/SohamMody/Event-Detection-from-Dockless-Scooter-Trips-Austin.

There is a notebook for each step of the project -- data preprocessing, exploratory data analysis, and modeling.

Appendix 1. Datasets descriptions

Datasets used

Data	Description	Data characteristics	Comments
Dockless Vehicle Trips (archived)	Archived version of the dataset above downloaded before geographical coordinates of were removed.	Date Range: April 2018 - February 2019 Table with numeric and text data. Columns include device id, trip duration, dates, council districts, census tract number Frequency: per scooter	Large amount of data (2.7 million rows) requires high processing power
Census Tract Fill shapefile	Spatial/Polygon representation of Austin at the following geographical levels of units: council districts, census tract	Polygons, boundaries are defined by the geographic representation level	

Datasets consulted

Data	Description	Data characteristics	Comments
Dockless Vehicle Trips	This dataset contains dockless vehicle trip data reported to the City of Austin Transportation Department as part of the Dockless Mobility Service operating rules.	Table with numeric and text data. Columns include device id, trip duration, dates, council districts, census tract number Frequency: per scooter Date Range: April 2018 - Present (May 2019)	Not used due to latitude and longitude data being removed in April 2019 and is aggregated on the council district level which is too broad.
Council District Fill shapefile	Spatial/Polygon representation of Austin at the following geographical levels of units: council districts, census tract	Polygons, boundaries are defined by the geographic representation level	This shapefile was used first, but was replaced by one with lower level of granularity (Census Tract fill shapefile)